The Predictive Performance of the Minnesota Screening Tool Assessing Recidivism Risk (MnSTARR): An External Validation

Authors
Grant Duwe, Ph.D.
Research Director
1450 Energy Park Drive, Suite 200
St. Paul, MN 55108-5219
Email: grant.duwe@state.mn.us

Michael Rocque, Ph.D.
Bates College
Department of Sociology
265 Pettengill Hall
Lewiston, Maine 04240
Email: mrocque@bates.edu

1450 Energy Park Drive, Suite 200
St. Paul, Minnesota 55108-5219
651/361-7200
TTY 800/627-3529
www.doc.state.mn.us
November 2019

This information will be made available in alternative format upon request.
Printed on recycled paper with at least 10 percent post-consumer waste

Research Summary

Using multiple performance metrics, this study externally validates the Minnesota Screening Tool Assessing Recidivism Risk (MnSTARR) among a sample of 3,985 inmates released from Minnesota prisons in 2014. While the Minnesota Department of Corrections implemented a fully-automated risk assessment (MnSTARR 2.0) in 2016, the original MnSTARR was a manually-scored, gender-specific recidivism risk assessment that predicted multiple types of recidivism—felony, non-violent, violent, and both first-time and repeat sexual offending (only for males). The results show the MnSTARR achieved adequate predictive performance. The average area under the curve (AUC) was 0.73 for males and 0.77 for females. Nonetheless, the MnSTARR would have achieved better predictive performance had it used an automated scoring process. Further, the findings showed the MnSTARR performed better for Whites than Non-Whites, and the magnitude of this difference would have been minimized using automated scoring. In sum, while the MnSTARR had adequate validity, performance is likely to be improved with automated systems.

Introduction

Correctional authorities use risk assessments to guide a host of decisions that are intended to not only make better use of scarce resources but also to enhance institutional and public safety. Risk assessment instruments have been used, for example, to help determine institutional custody levels for inmates, whether prisoners should be paroled, and the intensity of supervision for probationers and parolees (Cunningham & Sorensen, 2006; Meredith et al., 2007; Viglione et al., 2015). Because institutional and community programming resources are limited, risk assessments have been used to identify which persons to prioritize for programming. And, in the case of sex offenders, risk assessment instruments sometimes influence decisions relating to community notification and involuntary civil commitment.

To perform well in predicting the outcome, which is often recidivism for correctional populations, risk assessments must be reliable and valid. One critique that has been lodged against correctional risk assessment tools, however, is that reliability and validity get lost in the rush to "innovation" (Baird, 2009, 2). Within the context of risk assessment, reliability refers to consistency, either between the raters who manually score the assessment or how well the items on an instrument correlate with one another (DeVellis, 2012). Validity, on the other hand, generally refers to accuracy in making correct predictions, or a correlation between a risk score and the outcome. Reliability and validity are intertwined insofar as an unreliable instrument will, by necessity, have diminished validity (Duwe & Rocque, 2017; Jackson, 2012).

When risk assessments are created, there are three types of validity that are critical—apparent, internal, and external (Harrell et al., 1996). Apparent validity refers to the predictive performance on the sample used to develop an assessment, whereas internal validity examines the extent to which an assessment's accuracy can be reproduced on the population underlying the

sample. Several internal validation procedures have been developed to determine the reproducibility of a prediction model, including the split-population method, k-fold validation, and bootstrap resampling (Harrell, 2001; Steyerberg et al., 2001). Meanwhile, external validity examines the predictive performance of an instrument on a sample other than the one used to develop and internally validate it. Evaluating how well a risk assessment instrument has performed in practice on a correctional population is an example of an external validation.

Because classification algorithms, such as logistic regression, are often designed to maximize fit, overfitting is one of the most serious concerns involved with the creation of a prediction tool. That is, after a risk assessment has been trained on the development sample (or training set), overfitting produces a reduction or "shrinkage" in predictive accuracy when the tool is applied to another sample, such as the "test set" or "validation sample." Apparent validity provides an overly optimistic assessment of model performance given that it looks at an instrument's predictive accuracy on the development sample. Because internal and external validity assess predictive accuracy on non-development samples, both help determine the degree of optimism associated with apparent validity, which is reflected in the amount of shrinkage in accuracy when the tool is applied to other samples.

When recidivism risk assessment instruments are trained and tested, the developers of these tools seldom have the data available to conduct an external validation. Instead, the focus is almost invariably on internal validity (Duwe, 2014). Therefore, even though there is often "shrinkage" in predictive accuracy from the training set (or development sample) to the test set (or validation sample), the performance of an assessment on the test set still provides an estimate of how it will perform when it "goes live" and is used in practice. But when the assessment "goes live," what does its performance look like compared to its performance when it was

internally validated? Is there additional shrinkage when it is rolled out and used in the field? Put another way, to what extent is there variation in predictive performance from the internal validation to an external validation?

Although external validation studies are not as common as they ought to be in the correctional risk assessment literature, they are important for several reasons. First and foremost, an external validation is critical to determining whether an instrument has performed well in predicting the targeted outcome. Second, an external validation can help highlight whether actual use of the instrument is consistent with how it was designed to be used. In other words, is the tool being used to guide treatment, and is it reducing recidivism? Finally, an external validation provides an opportunity to examine how an instrument has performed across different sub-groups of the correctional population. Differential validity is a concern that has emerged with respect to actuarial risk assessment. That is, it is unclear whether certain tools work as well for males versus females or whites versus non-whites. To the extent that risk assessments have greater predictive validity for certain groups compared to others, their value is diminished as correctional tools.

Recent debate has arisen, for example, over the performance of risk assessment tools across gender, race and ethnicity (Holtfretter & Cupp, 2007; Smith, Cullen, & Latessa, 2009). Because risk assessments often draw heavily on criminal history, which is often more pronounced for racial and ethnic minorities (see, e.g., Everett & Wojtkiewicz, 2002), some have argued that race is not neutral with respect to risk assessment and the very process of assessing risk for recidivism is inherently biased (Harcourt, 2015). Moreover, for more subjective assessments, this concern may also be linked to attributional bias, whereby non-white individuals are judged as more dangerous than whites, resulting in higher risk scores (Bridges & Steen,

1998). Limited work has not, thus far, indicated significant bias in risk assessment (Skeem & Lowenkamp, 2016); however the question of whether actuarial risk assessment tools apply or are equally valid with all racial/ethnic groups is still open (McCafferty, 2018).

Present Study

In this study, we carry out an external validation of the Minnesota Screening Tool Assessing Recidivism Risk (MnSTARR), which the Minnesota Department of Corrections (MnDOC) began using in 2013 to assess recidivism risk for its prison population. Although the MnSTARR was internally validated when it was developed, we evaluate its predictive performance on the first cohort of inmates who were released from prison with a MnSTARR assessment. Overall, there were 3,985 inmates (3,585 males and 400 females) who had been assessed on the MnSTARR prior to their release from prison in 2014.

The MnDOC used the MnSTARR, which was manually scored by prison caseworkers, until it was replaced by the MnSTARR 2.0—a fully-automated assessment—in November 2016. Prior research suggests that an automated scoring process can improve predictive performance in comparison to a manual approach by eliminating inter-rater disagreement (Duwe & Rocque, 2017). Because the MnSTARR contains relatively objective items that can be pulled electronically from the same databases used for the MnSTARR 2.0, we examine whether a fully-automated scoring process would have affected predictive performance among the 3,985 prisoners who received a manual MnSTARR assessment prior to their release from prison.

Given recent concerns over bias in the design and use of risk assessment instruments, we also analyze the MnSTARR's performance by gender and race/ethnicity. In particular, for both males and females, we evaluate whether the MnSTARR differentially predicts recidivism for Whites compared to Non-Whites. Moreover, if there is a difference between Whites and non-

Whites, we further examine whether the use of an automated scoring method would have improved or exacerbated the difference.

In the next section, we begin by reviewing the state of risk assessment within corrections. After describing the development, validation and implementation of the MnSTARR, we discuss the data and methods used in this study. Following our presentation of the results from the MnSTARR external validation, we conclude by discussing the implications for risk assessment research and practice.

Recent Risk Assessment Research

As has now been widely documented, risk assessment in corrections has developed in several stages. First-generation assessment involved professionals using their experience to make decisions about which offenders were more likely to recidivate. Second-generation assessment introduced actuarial tools, which used quantifiable factors to assess risk in an effort to improve validity and reliability of assessments. Actuarial tools included risk factors that could, ideally, be scored objectively, with risk factors added together to arrive at an overall score. Research has consistently shown that actuarial tools have better validity than professional judgement (Abbott, 2011; Bonta & Andrews, 2007; Duwe & Rocque, 2018). Third-generation assessment built upon the success of second generation tools to include dynamic or changing factors and point to areas of possible intervention. Finally, fourth-generation tools expanded the scope of third generation tools to include more case management guidance (Bonta & Andrews, 2007). A recent study found that risk assessment in corrections is common in the United States, with at least 19 different tools having undergone some form of validity test (Desmarais, Johnson, & Singh, 2016).

Of the recent advances in risk assessment, one of the most prominent is the development of fully-automated risk assessment instruments. The process in which the items on a risk assessment instrument are populated has been referred to as the scoring method (Duwe & Rocque, 2017). The values for items can be entered manually, or they can be populated through an automated process. [1] When a manual scoring approach is used, differences in how staff score an assessment are often inevitable. For example, when staff manually score an assessment, they must interpret the information they obtain (either from a face-to-face interview with the individual or a database review), make decisions on what the appropriate response is for each item, and then correctly enter the values for these items on the instrument. Agreement among the staff who manually score the instrument can be difficult to achieve due to a variety of factors such as the subjectivity of the items on the instrument, inadequate training, staff workloads, the amount of time it takes to complete an assessment, and data entry errors.

Due to the idiosyncratic nature of hand-scoring instruments, concerns about reliability between raters have arisen. Some work has found non-trivial differences across scorers (Rocque & Plummer-Beale, 2014; Schmidt, Hoge, & Gomes, 2005; van der Knaap, Leenarts, Born, & Oosterveld, 2012), although overall reliability appears to be adequate. Still, an automated scoring process eliminates inter-rater disagreement by scoring each assessment the same way. And, given that reliability affects validity, the absence of inter-rater disagreement leads to better predictive performance (Duwe & Rocque, 2017).

One study evaluated the effect of automation on reliability and validity, showing that a fully-automated assessment is more reliable and valid (Duwe & Rocque, 2017). Duwe and

[1] The classification method is often synonymous with the scoring method. For example, the automated scoring process is frequently conflated with machine learning algorithms. However, it is possible to design and implement a fully-automated risk assessment that runs on a very simple Burgess algorithm. Conversely, it is also possible to create a manually-scored assessment that runs on a machine learning algorithm.

Rocque (2017) also found that an automated scoring process is more efficient and cost-effective. Compared to an assessment that must be scored manually, staff do not have to spend time scoring an automated assessment or undergo the significant amount of training that is required to maintain a manual tool. Moreover, the automation of the scoring process can produce a substantial increase in assessment capacity. Due to the savings in staff time, Duwe and Rocque (2017) reported that automation of the Minnesota Screening Tool Assessing Recidivism Risk (MnSTARR) 2.0—a recidivism risk assessment developed and validated on Minnesota's prisoner population—would yield a return on investment (ROI) of more than $20 after five years, generating close to $5 million in staff time saved.

In addition to automation and the impact it has on reliability and predictive validity, recent risk assessment research has examined home-grown assessments versus generic, off-the-shelf tools. The results from several studies lend credence to the notion that customized assessments may have a "home-field advantage" in achieving better predictive performance (Duwe & Rocque, 2018). Because local instruments can include factors relevant to local contexts overlooked by generic tools designed for broad use (Miller & Lin, 2007), existing research has shown that locally-developed risk assessment tools generally outperform those created and validated elsewhere (Drake, 2014). Most recently, Duwe and Rocque (2018) found that a home-grown instrument (the MnSOST-3) outperformed a global assessment (the Static-99R) in predicting sexual recidivism for 650 sex offenders released from Minnesota prisons in 2012.

While home-grown tools may outperform generic ones for specific populations, it is important to assess whether those tools apply to populations outside of those in which they were developed. For example, risk assessment tools often undergo initial validation assessments within development and test samples (See e.g., Duwe, 2014; Picard-Fritsche et al., 2017). These

steps provide information regarding the effectiveness of the tools in predicting recidivism. However, in order to determine whether the tool is effective "in practice," it must also be validated within the population for which it will be employed and for various groups within that population. This form of validity assessment is the basis of the current study.

Differential Validity and Risk Assessment

Another important issue in the evaluation of risk assessment tools is whether they apply equally across sub-groups. Certain early research pointed out the possibility of differential validity in the use of self-report surveys. Hindelang, Hirschi, and Weis (1981), for example, found that a self-report of delinquency instrument had higher validity among white respondents than among African-Americans. Limited scholarship has examined differential validity with respect to risk assessment tools. Some work has indicated that tools are less valid for females compared to males (Anderson et al., 2016) and whites or Hispanics compared to African-Americans (Rembert, Henderson, & Pirtle, 2013; Schwalbe et al., 2006). However, other work has argued that certain tools, despite some variation, are effective across groups (Harer & Langan, 2001; Smith, Cullen, & Latessa, 2009; Thompson & McGrath, 2012). In short, the question of differential validity with respect to risk assessment remains an important one that is deserving of more research. To the extent that tools are not equally valid, their value is diminished.

Development, Validation and Implementation of the MnSTARR

In April 2013, the MnDOC implemented the Minnesota Screening Tool Assessing Recidivism Risk MnSTARR, a "multiple-band" instrument that assesses risk separately for male and female prisoners for five different types of recidivism—nonviolent, felony, nonsexual violent, first-time sexual offending, and repeat sexual offending—over a 4-year follow-up period

(Duwe, 2014). The felony recidivism measure includes both violent and non-violent offenses, while the other four recidivism outcomes include felony, gross misdemeanor and, in some instances, misdemeanor offenses. By measuring the type and severity of reoffending, the five outcomes collectively provide a comprehensive assessment of recidivism risk.

The MnSTARR was developed on the assumption that risk factors vary by gender, resulting in separate recidivism risk scales for males and females (Duwe, 2014). Both males and females were assessed for their risk of nonviolent, felony, and nonsexual violent recidivism, although only males are assessed for their risk of either first-time or repeat sexual offending.[2] Males without a history of sexual offending were assessed for their risk of committing a first-time sex offense, whereas those with a sexual offending history were assessed for their risk of sexual recidivism.[3]

For the male version of the MnSTARR, the instrument contained either 23 items (non-sex offenders) or 24 (sex offenders). The female version of the instrument contained 19 items. Roughly three-fifths of the items on both the male and female versions of the MnSTARR pertained to criminal history. More specifically, the MnSTARR used detailed, comprehensive measures of criminal history that were disaggregated by type of offense and, in some instances, by the timing of the offense (i.e., how old the individual was at the time of the offense or how recently the offense took place).

Because the MnSTARR was created to be a risk assessment tool, it was not designed to identify which needs areas should be targeted for programming. Yet, because the MnSTARR's

[2] None of the 1,100 female offenders who made up the development sample for the female version of the MnSTARR recidivated with a new sex offense within 4 years of release from prison. As a result, the absence of female offenders who recidivated with a new sex offense made it impossible to develop a risk scale for either first-time or repeat sexual offending (Duwe, 2014)

[3] The risk scale for first-time sexual offending was derived from the development of the Minnesota Sexual Criminal Offending Risk Estimate (MnSCORE) (Duwe, 2012), whereas the sexual recidivism scale was drawn from the Minnesota Sex Offender Screening Tool-3 (MnSOST-3) (Duwe & Freske, 2012).

noncriminal history/dynamic items measured observable behavior in prison such as misconduct or completion of programming, the assessment indicated which needs areas improved or grew worse while an individual was incarcerated. For example, a major criminogenic need is antisocial peers. On the MnSTARR, active membership in a security threat group (i.e., gang) is a dynamic factor—because offenders can gain or lose active membership while in prison—that increases risk for male offenders. In contrast, receiving visits in prison, which generally increases prosocial support and has been associated with reduced recidivism in Minnesota (Duwe & Clark, 2013) and elsewhere (Bales & Mears, 2008), decreases risk for some measures of recidivism. Similarly, completing chemical dependency treatment in prison, which addresses substance abuse (a moderate criminogenic need), lowers an individual's recidivism risk according to the MnSTARR.

In the MnSTARR development study, the overall sample consisted of 11,375 males and 1,100 females who were released from prison between 2003 and 2006. Multiple logistic regression was the classification method used to develop the MnSTARR, and backward stepwise selection, along with bootstrap resampling, was used to identify significant predictors.[4] After estimating the final logistic regression models, which included interaction terms, for the eight recidivism measures (five for males and three for females) on the MnSTARR, Duwe (2014) used bootstrap resampling to generate optimism-corrected estimates of predictive performance.

Using the AUC as the lone metric for predictive validity, Duwe (2014) reported the optimism-corrected estimates ranged from 0.73 to 0.80 across the five recidivism measures for male offenders and from 0.73 to 0.81 for the three recidivism measures for female offenders. Because the MnSTARR was originally designed to be a manually-scored instrument via a

[4] Because backward selection is generally preferable to forward selection (Harrell et al., 1996), it was the approach Duwe (2014) used for the MnSTARR.

database review, the development study also included an inter-rater reliability assessment among MnDOC caseworkers that found an overall intraclass correlation coefficient (ICC) of 0.84 for the eight recidivism measures (Duwe, 2014).

Prior to the MnSTARR development study, the MnDOC had used the Level of Service/Case Management Inventory (LS/CMI) and, before that, the Level of Service Inventory-Revised (LSI-R) to assess risk and need. But given that the MnSTARR significantly outperformed the LSI-R in predicting multiple types of recidivism for Minnesota prisoners, the MnDOC began using the MnSTARR as its risk assessment instrument in April 2013. Per MnDOC policy, administration of the MnSTARR was limited to prisoners who were confined more than 180 days. Thus, inmates whose imprisonment periods were less than 180 days did not receive a MnSTARR or LS/CMI assessment.

Similar to other fourth-generation assessment instruments, the MnSTARR was designed to be administered at least twice on a single person—once at the time of intake to help prioritize the higher-risk prisoners for institutional programming and one more time prior to release to help inform decisions relating to community supervision and programming. The MnDOC has continued to use the LS/CMI but strictly as a needs assessment instrument. More specifically, use of the LS/CMI has been limited to the high- and very high-risk (i.e., top 40%) offenders, per the MnSTARR, because these are the inmates who generally get prioritized for institutional programming.

In November 2016, the MnDOC transitioned from the MnSTARR—an assessment manually scored by correctional staff—to the MnSTARR 2.0—a fully-automated assessment (Duwe & Rocque, 2017). The MnSTARR 2.0 extracts data from the state's criminal history repository to populate the criminal history items on the instrument, while data from the

Correctional Operations Management System (COMS)—the MnDOC's centralized database—are pulled to populate items pertaining to demographic characteristics (e.g., gender, age, and marital status), institutional behavior (e.g., discipline convictions and gang affiliation), and participation in programming (e.g., earning a post-secondary degree in prison, completing chemical dependency treatment, and completing cognitive-behavioral therapy). The only MnSTARR 2.0 items that are not auto-populated are those for the MnSOST-3, which continued to be scored manually by correctional staff. Still, after a MnSOST-3 assessment has been completed, the MnSOST-3 score is extracted from COMS and uploaded within the MnSTARR 2.0 assessment.

Data and Method

Our sample consists of 3,985 (3,585 males and 400 females) inmates who had been assessed on the MnSTARR prior to their release from Minnesota prisons in 2014. As noted above, MnDOC policy dictated that only inmates who were going to be in prison for 180 days or more should be assessed on the MnSTARR. Of the 7,657 releases from Minnesota prisons in 2014, there were 4,392 (57%) who had a length of stay of six months or more. As a result, while 3,985 releases received a MnSTARR assessment, there were 407 who should have been assessed on the MnSTARR, but were not, prior to their release. Moreover, even though the MnSTARR was designed to be administered twice to each eligible inmate, only about one-third of those who had been assessed had multiple MnSTARR assessments.

To assess the MnSTARR's predictive performance, we focused on the last assessment prior to release for inmates who had more than one assessment. In addition, we obtained statewide reconviction data electronically from the Minnesota Bureau of Criminal Apprehension (BCA) to measure recidivism among the 3,985 releases. Because the MnSTARR was designed to

assess recidivism risk over a four-year follow-up period, we collected BCA reconviction data through the end of 2018 to ensure that everyone in our 2014 release cohort had a full four-year follow-up period.

Consistent with the development of the MnSTARR (Duwe, 2014), non-sexual violent reconvictions included all person crimes, except for sex offenses, regardless of severity level (misdemeanor, gross misdemeanor, and felony). Non-violent reconvictions contained all non-person crimes regardless of severity level. Felony reconvictions included all felony-level offenses regardless of the type of offense. Sex offense reconvictions included only hands-on sex offenses, and this measure was the same regardless of whether it was a first-time or repeat sex offense.

Predictive Performance Metrics

Similar to recent risk assessment studies that have used multiple statistics to evaluate predictive performance (Duwe, 2017; Duwe & Kim, 2016; Duwe & Rocque, 2017; Hamilton, Neuilly, Lee, & Barnoski, 2015; Tollenaar & van der Heijden, 2013), we used four different metrics to assess the three key dimensions of predictive validity—accuracy, discrimination, and calibration. The accuracy (ACC) statistic is one of the more commonly-used metrics for predictive accuracy, which assesses how well a model makes correct classification decisions. If an individual who recidivated had a predicted probability less than 50 %, then this person would be incorrectly classified (i.e., false negative). Conversely, if this individual did not recidivate, then s/he would be accurately classified (i.e., true negative). The ACC value ranges from 0 to 100%, and higher ACC values reflect greater accuracy in making correct classification decisions.

Predictive discrimination measures the degree to which an assessment separates—in this instance—the recidivists from those who do not recidivate within the follow-up window. One of

the most-widely used predictive discrimination statistics is the area under the curve (AUC), which is relatively robust across different recidivism base rates and selection ratios (Smith, 1996). The AUC statistic is interpreted as the probability that a randomly selected recidivist has a higher score on a risk assessment instrument than a randomly selected non-recidivist. According to the literature, an AUC between 0.90 and 1.00 is considered excellent, between 0.80 and 0.89 is good, between 0.70 and 0.79 is fair, between 0.60 and 0.69 is poor, and between 0.50 and 0.59 represents a failure to achieve predictive discrimination (Baird et al, 2013; Thornton & Laws, 2009).

Calibration looks at how well the predicted probabilities from a model correspond with the observed outcome being predicted. Calibration is, therefore, a measure of absolute risk, while predictive discrimination assesses relative risk. In order for a prediction instrument to make accurate absolute assessments of risk, the model's predicted probabilities must be calibrated with the observed recidivism outcomes. With values that range from 0 to 1, root mean square error (RMSE) measures the squared root of the average squared difference between observed recidivism and predicted probabilities. The closer the RMSE value is to zero, the better the calibration.

In addition to these metrics, we used a consolidated statistic to assess overall predictive performance. The SAR (squared error, accuracy, ROC (receiver operating characteristic)) is a combined measure of discrimination, accuracy and calibration, and its formula is: (ACC + AUC + (1 – RMSE))/3 (Caruana, Niculescu-Mizil, Crew, & Ksikes, 2004). In previous correctional research that has used the SAR, values have ranged from a low of 0.62 to a high of 0.90 (Duwe & Kim, 2016; Duwe & Rocque, 2017; Hamilton et al., 2015; Tollenaar & van der Heijden, 2013).

Results

The results in Table 1 show the observed recidivism rates in comparison to the average predicted probabilities. The findings suggest the MnSTARR overestimated risk for most of the recidivism outcomes predicted. The lone exception is violent recidivism for males, where the average predicted probabilities were similar to the actual recidivism rate. The results also indicate the manual scoring method overestimated risk more than the automated process.

Table 1. Average Predicted Probabilities and Observed Recidivism Comparison

	Recidivism Outcomes				
Males	Felony	Non-Violent	Violent	First-Time	Repeat
Observed	42.6	53.3	23.3	0.8	1.2
Predicted					
Manual	53.0	66.7	23.6	3.0	4.8
Automated	47.0	62.2	23.4	2.9	
Females					
Observed	35.3	52.5	6.8		
Predicted					
Manual	40.6	57.7	7.9		
Automated	38.1	52.9	7.1		

Table 2 shows the results for males across the five recidivism measures for each of the four predictive performance metrics. More specifically, the AUC results are presented for both apparent and optimism-corrected predictive validity for all five recidivism measures at the time these assessments were developed and validated. The "Manual" results reflect the predictive performance of the MnSTARR scored manually by MnDOC caseworkers. The "Automated" results, on the other hand, show how the MnSTARR would have performed had it been scored by an automated process. Because the MnSOST-3 contains some items that cannot be scored through an automated process, the repeat sex offending measure ("Repeat") does not include any "Automated" results.

Table 2. Predictive Performance Results: Males

	Felony	*Non-Violent*	*Violent*	*First-Time*	*Average*	*Repeat*	*Average*
AUC							
Apparent	0.785	0.772	0.758	0.818	0.783	0.821	0.791
Optimism-Corrected (O-C)	0.764	0.752	0.730	0.763	0.752	0.796	0.761
Apparent and O-C Difference	-0.019	-0.020	-0.028	-0.055	-0.031	-0.025	-0.030
Manual	0.714	0.733	0.747	0.662	0.714	0.778	0.727
Manual and O-C Difference	-0.050	-0.019	0.017	-0.101	-0.038	-0.018	-0.034
Automated	0.725	0.741	0.760	0.712	0.735		
Manual and Automated Difference	0.011	0.008	0.013	0.050	0.021		
Automated and O-C Difference	-0.039	-0.011	0.030	-0.051	-0.017		
ACC							
Manual	0.647	0.650	0.775	0.985	0.764	0.978	0.807
Automated	0.667	0.663	0.772	0.984	0.772		
Difference	0.020	0.013	-0.003	-0.001	0.008		
RMSE							
Manual	0.477	0.477	0.396	0.118	0.367	0.136	0.321
Automated	0.463	0.464	0.395	0.119	0.360		
Difference	-0.014	-0.013	-0.001	0.001	-0.007		
SAR							
Manual	0.628	0.635	0.709	0.843	0.704	0.873	0.738
Automated	0.643	0.647	0.712	0.859	0.715		
Difference	0.015	0.011	0.004	0.016	0.011		

AUC = Area under the Curve
ACC = Accuracy
RMSE = Root Mean Squared Error
SAR = Squared Error, Accuracy, Receiver Operating Characteristic

As displayed in Table 2, the average optimism-corrected AUC (0.761) was .03 lower than the average apparent AUC (0.791). When the MnSTARR was scored in practice by caseworkers, the average AUC (0.731) was .03 lower than the optimism-corrected AUC, which provided an estimate of how the MnSTARR might perform. The AUC for the manual assessments was lower than the optimism-corrected AUC for all of the recidivism measures except for violent reoffending. In particular, for first-time sexual offending, the AUC for the manual process (0.662) was .101 lower than the optimism-corrected AUC (0.763).

As noted earlier, a recent study reported an AUC of 0.716 for the MnSOST-3 (i.e., the sex offense recidivism assessment integrated within the MnSTARR) (Duwe, 2017; Duwe & Rocque, 2018). That study evaluated the MnSOST-3's predictive performance for a cohort of releases from prison in 2012. This study, however, found an AUC of 0.778, which is closer to the optimism-corrected value of 0.796, for sex offenders released from prison in 2014.

Table 3. Predictive Performance Results: Females

	Felony	*Non-Violent*	*Violent*	*Average*
AUC				
Apparent	0.743	0.765	0.819	0.776
Optimism-Corrected (O-C)	0.731	0.757	0.805	0.764
Apparent and O-C Difference	-0.012	-0.008	-0.014	-0.013
Manual	0.702	0.749	0.843	0.765
Manual and O-C Difference	-0.029	-0.008	0.038	0.001
Automated	0.709	0.760	0.858	0.776
Manual and Automated Difference	0.007	0.011	0.015	0.011
Automated and O-C Difference	-0.022	0.003	0.053	0.012
ACC				
Manual	0.665	0.693	0.928	0.762
Automated	0.665	0.710	0.933	0.769
Difference	0.000	0.017	0.005	0.007
RMSE				
Manual	0.462	0.458	0.241	0.387
Automated	0.458	0.450	0.228	0.379
Difference	-0.004	-0.008	-0.013	-0.008
SAR				
Manual	0.635	0.661	0.843	0.713
Automated	0.639	0.673	0.854	0.722
Difference	0.004	0.012	0.011	0.009

AUC = Area under the Curve
ACC = Accuracy
RMSE = Root Mean Squared Error
SAR = Squared Error, Accuracy, Receiver Operating Characteristic

Similar to Table 2, we present the predictive performance results for females in Table 3 across the three recidivism measures for each of the four predictive performance metrics. Unlike Table 2, however, we do not show results for either first-time or repeat sex offending considering

the MnSTARR does not contain these measures for females. At the time the MnSTARR was developed and validated, the average optimism-corrected AUC (0.764) was .013 lower than the average apparent AUC. The results in Table 3 show little overall difference from the optimism-corrected AUC (0.764) to the average AUC for manual assessment (0.765). Among the three

Table 4. Male Predictive Performance Results by Race/Ethnicity

	Felony	*Non-Violent*	*Violent*	*First-Time*	*Average*	*Repeat*	*Average*
AUC							
Non-White Manual	0.707	0.714	0.717	0.649	0.697	0.756	0.707
White Manual	0.720	0.752	0.776	0.637	0.721	0.827	0.739
Difference—Manual	0.013	0.038	0.059	-0.012	0.024	0.071	0.032
Non-White Automated	0.725	0.724	0.736	0.704	0.722		
White Automated	0.725	0.758	0.776	0.671	0.733		
Difference-Automated	0.000	0.034	0.040	-0.033	0.011		
ACC							
Non-White Manual	0.652	0.638	0.737	0.977	0.751	0.973	0.788
White Manual	0.641	0.664	0.815	0.993	0.778	0.982	0.812
Difference—Manual	-0.011	0.026	0.078	0.016	0.027	0.009	0.024
Non-White Automated	0.670	0.651	0.731	0.977	0.757		
White Automated	0.663	0.675	0.815	0.991	0.786		
Difference-Automated	-0.007	0.024	0.084	0.014	0.029		
RMSE							
Non-White Manual	0.478	0.482	0.426	0.144	0.383	0.147	0.357
White Manual	0.522	0.490	0.371	0.082	0.366	0.126	0.342
Difference—Manual	0.044	0.008	-0.055	-0.062	-0.017	-0.021	-0.015
Non-White Automated	0.463	0.470	0.426	0.144	0.376		
White Automated	0.507	0.476	0.369	0.084	0.359		
Difference-Automated	0.044	0.006	-0.057	-0.060	-0.017		
SAR							
Non-White Manual	0.627	0.623	0.676	0.827	0.688	0.861	0.717
White Manual	0.613	0.642	0.740	0.849	0.711	0.894	0.742
Difference—Manual	-0.014	0.019	0.064	0.022	0.023	0.033	0.025
Non-White Automated	0.644	0.635	0.680	0.846	0.701		
White Automated	0.627	0.652	0.741	0.859	0.720		
Difference-Automated	-0.017	0.017	0.059	0.013	0.019		

AUC = Area under the Curve
ACC = Accuracy
RMSE = Root Mean Squared Error
SAR = Squared Error, Accuracy, Receiver Operating Characteristic

types of recidivism, the AUC for the manual assessment was higher for violent recidivism but lower for felony and non-violent recidivism.

Table 5. Female Predictive Performance Results by Race/Ethnicity

	Felony	*Non-Violent*	*Violent*	*Average*
AUC				
Non-White Manual	0.716	0.731	0.776	0.741
White Manual	0.689	0.755	0.852	0.765
Difference—Manual	-0.027	0.024	0.076	0.024
Non-White Automated	0.721	0.749	0.838	0.769
White Automated	0.700	0.761	0.833	0.765
Difference-Automated	-0.019	0.012	-0.005	-0.004
ACC				
Non-White Manual	0.648	0.703	0.855	0.735
White Manual	0.675	0.686	0.969	0.777
Difference—Manual	0.027	-0.017	0.114	0.042
Non-White Automated	0.648	0.731	0.876	0.752
White Automated	0.675	0.698	0.965	0.779
Difference-Automated	0.027	-0.033	0.089	0.027
RMSE				
Non-White Manual	0.469	0.462	0.326	0.419
White Manual	0.457	0.456	0.175	0.363
Difference—Manual	-0.012	-0.006	-0.151	-0.056
Non-White Automated	0.466	0.454	0.301	0.407
White Automated	0.453	0.449	0.173	0.358
Difference-Automated	-0.013	-0.005	-0.128	-0.049
SAR				
Non-White Manual	0.632	0.657	0.768	0.686
White Manual	0.636	0.662	0.882	0.726
Difference—Manual	0.004	0.005	0.114	0.040
Non-White Automated	0.634	0.675	0.804	0.705
White Automated	0.641	0.670	0.875	0.729
Difference-Automated	0.007	-0.005	0.071	0.024

AUC = Area under the Curve
ACC = Accuracy
RMSE = Root Mean Squared Error
SAR = Squared Error, Accuracy, Receiver Operating Characteristic

When we compare the findings for the two types of scoring methods evaluated, we see a modest advantage in performance for the automated process. Overall, the AUC was .011 higher,

the ACC was .007 higher and the RMSE was .008 lower. As a result, the SAR for the automated method (0.722) was .009 higher than the manual process (0.713).

In Tables 4 and 5, we examine the predictive performance results by race and ethnicity by comparing White and Non-White for both males and females. In Table 4, we see better predictive performance results for White males in comparison to Non-Whites. For example, the average AUC was .024 higher (.032 higher when including sex offense recidivism), the average ACC was .027 higher, the average RMSE was .017 lower, and the average SAR was .023 higher. Automation minimized the difference, however, between Whites and Non-Whites, at least for the AUC. Whereas the average AUC was .024 higher for Whites with the manual methods, it was .011 higher when the automated process was used.

The results for females, which are depicted in Table 5, also show the manually-scored MnSTARR had better predictive performance for Whites. Indeed, compared to Non-White females, the average AUC for Whites was .024 higher, the ACC was .042 higher, the RMSE was .056 lower, and the SAR was .040 higher. Similar to the results for males, however, use of an automated process would have minimized the predictive performance differences between Whites and Non-Whites. Most notably, the average AUC for the automated process (0.769) for Non-Whites was actually .004 higher than the average for Whites (0.765). In addition, the differences between Whites and Non-Whites for the other three metrics (ACC, RMSE and SAR) were smaller for the automated process in comparison to the manual approach.

Conclusion

When developed, risk assessment tools often go through a series of validation tests, moving from internal to external samples. These tests, including examinations of applicability of instruments across subgroups, are essential in determining how useful particular tools are for

improving public safety. This study conducted an external validity assessment of the MnSTARR using a sample of offenders released in 2014 from the MnDOC. It also examined differential validity across sex and racial groups.

The findings indicate the MnSTARR achieved adequate predictive performance among the first cohort of releases from prison in 2014. For females, the average AUC was 0.765 and the SAR was a respectable 0.713. For males, the average AUC across all five recidivism measures was 0.727 and the average SAR was 0.738. For both males and females, predictive performance was best for violent recidivism, followed by non-violent and then felony recidivism. For males, predictive performance was better for sex offense recidivism than it was for first-time sexual offending.

The optimism-corrected AUC value provided a better estimate of predictive discrimination for females than it did for males. The average AUC from the external validation of the MnSTARR for females (0.765) was off by only .001 than the average optimism-corrected AUC (0.764). In particular, the AUC was higher for the external validation than it was for the optimism-corrected estimate for violent recidivism, while the opposite was true for both felony and, to a lesser extent, non-violent recidivism.

For males, the average AUC from the external validation was .034 lower the optimism-corrected estimates. Therefore, at least for males, the optimism-corrected AUC values from the internal validation were still too optimistic. Much of the difference between the internal and external validation AUC values was due to the felony and first-time sex offending measures. In fact, the external validation AUC was .101 lower than the optimism-corrected estimate for first-time sex offending. Of course, if an automated scoring approach had been used, then the difference would have been smaller (.051).

More generally, consistent with the prior study by Duwe and Rocque (2017), the findings indicate the use of an automated scoring method would have produced a modest benefit in predictive performance, increasing the AUC, on average, by .011 for females and .021 for males. In addition to better predictive performance overall, the results suggest an automated scoring process would have improved performance more for Non-Whites than Whites. Despite achieving adequate predictive validity for both Whites and Non-Whites, the MnSTARR still performed better for Whites than Non-Whites. The use of an automated scoring process, however, would have helped minimize this difference between Whites and Non-Whites.

These results should not be taken to mean that staff were biased (intentionally or otherwise) in how they scored the MnSTARR. Instead, these findings likely demonstrate the benefits of increased reliability. When inter-rater disagreement is present, which is virtually inevitable for manual assessments, it may increase the possibility that group differences will emerge in predictive performance. Conversely, the likelihood of group differences may be diminished when every inmate is scored the same way on an assessment, which is true for an automated process.

While the findings from this study suggest an automated scoring process can yield improvements in predictive performance, it is also worth underscoring the impact it can have on assessment capacity. Only 52 percent of the 7,657 releases from Minnesota prisons in 2014 had been assessed for recidivism risk, and most of those assessed had received only one assessment (instead of at least two). Put another way, nearly half of all releases from prison had not been assessed for risk, due in part to the amount of time it takes to complete a manual assessment.

In contrast, since the MnSTARR 2.0 was implemented in November 2016, every individual released from Minnesota prisons has been assessed at least once and, in most

instances, multiple times prior to release. In our prior study on the MnSTARR 2.0 (Duwe and Rocque, 2017), we estimated that automating the scoring process would produce approximately 22,200 assessments per year. Because it took MnDOC staff, on average, 35 minutes to manually score a MnSTARR assessment, we estimated that 22,200 assessments would require 12,950 staff hours. Therefore, by saving that many hours of staff time, automation of the MnSTARR would yield a cost-benefit estimate of $452,108, resulting in a ROI of $4.35.

As it turns out, however, we underestimated the impact that automation would have on assessment capacity. Following its implementation in November 2016, a total of 41,253 MnSTARR 2.0 assessments were completed during the first year. If we again assume it would have taken 35 minutes, on average, to manually score this many assessments, automating the scoring process saved more than 24,000 hours in staff time (nearly the equivalent of 12 full-time employees), resulting in a revised cost-benefit estimate of $955,990 and a ROI of $8.08.

Limitations and suggestions for future research

This study has a few limitations that should be kept in mind for future research. First, the racial and ethnic comparisons used were limited to Whites vs. Non-whites due to minimal variation in racial/ethnic sample size. The MnSTARR may work better for certain Non-White groups compared to others. Second, the measures of reoffending were limited to officially-recorded criminal behaviors rather than self-reported activity. It is possible that the findings may have differed for unreported offenses. Future research should seek to examine validity of risk assessment instruments with self-reported offending and perhaps compare the results to those with more common officially recorded (e.g., convictions) measures.

As we alluded to earlier, the risk assessment literature would also benefit from the publication of more external validation studies that assess predictive performance on assessments

that have been completed in practice. In general, correctional agencies either implement "off-the-shelf" risk assessment instruments that have been developed and, in some cases, validated on other correctional populations, or they implement customized assessments (like the MnSTARR) that were developed and validated on their own population. In either case, carrying out an external validation on actual assessments is critical towards helping identify promising and proven practices in the creation and implementation of risk assessment instruments.

REFERENCES

Abbott, B. R. (2011). Throwing the baby out with the bath water: is it time for clinical judgment to supplement actuarial risk assessment? *Journal of the American Academy of Psychiatry and the Law Online*, *39*(2), 222-230.

Anderson, V. R., Davidson, W. S., Barnes, A. R., Campbell, C. A., Petersen, J. L., & Onifade, E. (2016). The differential predictive validity of the Youth Level of Service/Case Management Inventory: The role of gender. *Psychology, Crime & Law*, *22*, 666-677.

Baird, C. (2009). A question of evidence: A critique of risk assessment models used in the justice system. *Madison, WI: National Council on Crime and Delinquency*.

Baird, C., Healy, T., Johnson, K., Bogie, A., Dankert, E.W., & Scharenbroch, C. (2013). A comparison of risk assessment instruments in juvenile justice. National Council on Crime & Delinquency. Retrieved from https://www.nccdglobal.org/sites/default/files/publication_pdf/nccd_fire_report.pdf

Bales, W.D. & Mears, D.P. (2008). Inmate social ties and the transition to society: Does visitation reduce recidivism? *Journal of Research in Crime & Delinquency, 45*, 287–321.

Bonta, J. & Andrews, D.A. (2007). Risk-Needs-Responsivity Model for Offender Assessment and Rehabilitation. Ottawa: Public Safety Canada.

Bridges, G. S., & Steen, S. (1998). Racial disparities in official assessments of juvenile offenders: Attributional stereotypes as mediating mechanisms. *American Sociological Review*, *63*, 554-570.

Caruana, R., Niculescu-Mizil, A., Crew, G., & Ksikes, A. (2004). Ensemble selection from libraries of models, in *Proceedings of the 21st International Conference on Machine Learning*, Canada: Banff, 1-12.

Cunningham, M. D., & Sorensen, J. R. (2006). Actuarial models for assessing prison violence risk: revisions and extensions of the risk assessment scale for prison (RASP). *Assessment*, *13*(3), 253-265.

Desmarais, S. L., Johnson, K. L., & Singh, J. P. (2016). Performance of recidivism risk assessment instruments in U.S. correctional settings. *Psychological Services, 13*(3), 206-222.

DeVellis, R.F. (2012). Scale development: Theory and application. Newbury Park, CA: Sage.

Drake, E. (2014). *Predicting criminal recidivism: A systematic review of offender risk assessments in Washington State* (Doc. No. 14-02-1901). Olympia: Washington State Institute for Public Policy.

Duwe, G. (2012). Predicting first-time sexual offending among prisoners without a prior sex offense history: The Minnesota Sexual Criminal Offending Risk Estimate (MnSCORE). *Criminal Justice and Behavior*, *39*, 1,434-1,454.

Duwe, G. (2014). The development, validity, and reliability of the Minnesota Screening Tool Assessing Recidivism Risk (MnSTARR). *Criminal Justice Policy Review*, *25*, 579-613.

Duwe, G. (2017). Better practices in the development and validation of recidivism risk assessments: The Minnesota Sex Offender Screening Tool-4. *Criminal Justice Policy Review*. https://doi.org/10.1177%2F0887403417718608.

Duwe, G., & Clark, V. (2013). Blessed be the social tie that binds the effects of prison visitation on offender recidivism. *Criminal Justice Policy Review*, *24,* 271-296.

Duwe, G. & Freske, P. (2012). Using logistic regression modeling to predict sex offense recidivism: The Minnesota Sex Offender Screening Tool-3 (MnSOST-3). *Sexual Abuse: A Journal of Research and Treatment*, *24*, 350-377.

Duwe, G. & Kim, K. (2016). Sacrificing accuracy for transparency in recidivism risk assessment: The impact of classification method on predictive performance. *Corrections: Policy, Practice and Research*, *1*, 155-176.

Duwe, G. & Rocque, M. (2017). The effects of automating recidivism risk assessment on reliability, predictive validity, and return on investment (ROI). *Criminology & Public Policy*, *16*, 235-269.

Duwe, G. & Rocque, M. (2018). The home-field advantage and the perils of professional judgment: Evaluating the performance of the Static-99R and the MnSOST-3 in predicting sexual recidivism. *Law and Human Behavior*, *42*, 269-279.

Everett, R.S. & Wojtkiewicz, R.A. (2002). Difference, disparity, and race/ethnic bias in Federal sentencing. *Journal of Quantitative Criminology*, *18*, 189-211.

Hamilton, Z., Neuilly, M-A., Lee, S., & Barnoski, R. (2015). Isolating modeling effects in offender risk assessment. *Journal of Experimental Criminology*, *11*(2), 299–318.

Harcourt, B.E. (2015). Risk as a proxy for race: The dangers of risk assessment. *Federal Sentencing Reporter*, *27*, 237-243.

Harer, M. D., & Langan, N. P. (2001). Gender differences in predictors of prison violence: Assessing the predictive validity of a risk classification system. *Crime & Delinquency, 47*(4), 513-536.

Harrell, F.E. (2001). *Regression Modeling Strategies with Application to Linear Models, Logistic Regression, and Survival Analysis*. New York: Springer Verlag.

Harrell, F.E., Lee, K.L., & Mark, D. B. (1996). Multivariable prognostic models: Issues in developing models, evaluating assumptions and adequacy, and measuring and reducing errors. *Statistics in Medicine*, 15, 361-387.

Hindelang, M.J., Hirschi, T., & Weis, J.G. (1981). *Measuring Delinquency*. Thousand Oaks, CA: Sage.

Holtfretter, K., & Cupp, R. (2007). Gender and risk assessment: the empirical status of the LSI-R for women. *Journal of Contemporary Criminal Justice*, *23*, 363-382.

Jackson, S. L. (2012). *Research Methods: A Modular Approach*. Stamford, CT: Cengage.

McCafferty, J. T. (2018). Professional discretion and the predictive validity of a juvenile risk assessment instrument: Exploring the overlooked principle of effective correctional classification. *Youth Violence and Juvenile Justice*, *15*(2), 103-118.

Meredith, T., Speir, J. C., & Johnson, S. (2007). Developing and implementing automated risk assessments in parole. *Justice Research and Policy*, *9*(1), 1-24.

Miller, J., & Lin, J. (2007). Applying a generic juvenile risk assessment instrument to a local context some practical and theoretical lessons. *Crime & Delinquency*, *53*(4), 552-580.

Picard-Fritsche, S., Rempel, M., Tallon, J.A., Adler, J., & Reyes, N. (2017). *Demystifying risk assessment: Key principles and controversies*. Center for Court Innovation: New York.

Rembert, D.A., Henderson, H., & Pirtle, D. (2014). Differential racial/ethnic predictive validity. *Youth Violence and Juvenile Justice*, *12*, 152-166.

Rocque, M., & Plummer-Beale, J. (2014). In the eye of the beholder? An examination of the inter-rater reliability of the LSI-R and YLS/CMI in a correctional agency. *Journal of Criminal Justice*, *42*, 568-578.

Schmidt, F., Hoge, R. D., & Gomes, L. (2005). Reliability and Validity Analyses of the Youth Level of Service/Case Management Inventory. *Criminal Justice and Behavior, 32*(3), 329-344.

Schwalbe, C., Fraser, M., Day, S., & Cooley, V. (2006). Classifying juvenile offenders according to risk of recidivism: Predictive validity, race/ethnicity, and gender. *Criminal Justice and Behavior*, *33*, 305–24.

Skeem, J. L., & Lowenkamp, C. T. (2016). Risk, race, and recidivism: Predictive bias and disparate impact. *Criminology*, *54*, 680-712.

Smith, W. (1996). The effects of base rate and cutoff point choice on commonly used measures of association and accuracy in recidivism research. *Journal of Quantitative Criminology*, *12*, 83-111.

Smith, P., Cullen, F.T., & Latessa, E.J. (2009). Can 14,737 women be wrong? A meta-analysis of the LSI-R and recidivism for female offenders. *Criminology & Public Policy*, *8*, 183-208.

Steyerberg, E.W., Harrell, F.E., Borsboom, G.J.J.M., Eijkemans, M.J.C., Vergouwe, Y., & Habbema, J.D.F. (2001). Internal validation of predictive models: Efficiency of some procedures for logistic regression analysis. *Journal of Clinical Epidemiology*, 54, 774-781.

Thompson, A. P., & McGrath, A. (2012). Subgroup differences and implications for contemporary risk-need assessment with juvenile offenders. *Law and Human Behavior, 36*, 345-355.

Thornton, D., & Laws, D. R. (2009). *Cognitive Approaches to the Assessment of Sexual Interest in Sexual Offenders*. Hoboken, NJ: Wiley.

Tollenaar, N., & van der Heijden, P.G.M. (2013). Which method predicts recidivism best? A comparison of statistical, machine learning and data mining predictive methods. *Journal of the Royal Statistical Society*, Series A 176 (part 2): 565-584.

van der Knapp, L., Born, M.P., Leenarts, L.E.W., & Oosterveld, P. (2012). Reevaluation inter-rater reliability in offender risk assessment. *Crime & Delinquency*, *58*, 147-163.

Viglione, J., Rudes, D. S., & Taxman, F. S. (2015). Misalignment in supervision: Implementing risk/needs assessment instruments in probation. *Criminal Justice and Behavior*, *42*(3), 263-285.